DOCUMENT RESUME

ED 463 294 TM 033 717

AUTHOR Burroughs, Monte
TITLE Bootstrapping Selected Item Statistics from a Student-Made Test.
PUB DATE 2002-02-00
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (25th, Austin, TX, February 14-16, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Graduate Students; Graduate Study; Higher Education; *Hypothesis Testing; *Nonparametric Statistics; Statistical Analysis; *Test Items
IDENTIFIERS *Bootstrap Methods; Discrimination Parameters; Null Hypothesis

ABSTRACT

This study applied nonparametric bootstrapping to test null hypotheses for selected statistics (KR-20, difficulty, and discrimination) derived from a student-made test. The test, administered to 21 students enrolled in a graduate-level educational assessment class, contained 42 items, 33 of which were analyzed. Random permutations of the data yielded a bootstrapped mean KR-20 equal to 0.733 (p=0.012), a bootstrapped mean level of difficulty equal to 0.769 (p=0.212), and a bootstrapped mean point-biserial correlation equal to 0.302 (p=0.273). The bootstrapped KR-20 was unusual given random permutations of the data. Results failed to reject the other two null hypotheses, suggesting each model was not unusual given random permutations of the data. (Contains 3 figures and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 294

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Burroughs

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Bootstrapping selected item statistics from a student-made test[1]

Monte Burroughs

University of North Carolina at Greensboro

TM033717

[1] A paper presented at the annual meeting of the Southwest Educational Research Association, Austin, Texas, February 14 - February 16, 2002.

BEST COPY AVAILABLE

Summary

This study applied nonparametric bootstrapping to test null hypotheses for selected statistics (KR-20, difficulty, and discrimination) derived from a student-made test. The test, administered to 21 students (n = 21) enrolled in a graduate-level educational assessment class, contained 42 items, 33 of which were analyzed.

Random permutations of the data yielded a bootstrapped mean KR-20 equal to 0.733 (p = 0.012), a bootstrapped mean level of difficulty (DIFF) equal to 0.769 (p = 0.212), and a bootstrapped mean point-biserial correlation (PBIS) equal to 0.302 (p = 0.273). The bootstrapped KR-20 was unusual given random permutations of the data. Results failed to reject the other two null hypotheses, suggesting each model was not unusual given random permutations of the data.

Introduction

Writing a classroom test that can be scored objectively is more difficult than one may imagine. Students enrolled in teacher education programs spend considerable time learning about and writing lesson plans, unit objectives, and other theories associated with curriculum and instruction. Unfortunately, students receive scant instruction about how to construct useful and informative classroom tests or assessments, and receive even less, if any, instruction on test analysis. Yet, testing and assessment are integral components of teaching and learning.

This paper evolved from a two-part exercise designed to give students enrolled in teacher education a foretaste of issues and procedures associated with test construction at the classroom-level. First, I told the students that they would write their own mid-term exam. I randomly assigned students to groups and then randomly assigned each group a chapter from the course textbook. Every group had to write 15 questions in any combination of three formats (two-choice, multiple-choice, or short answer) that aligned with its respective chapter's objectives. I concatenated items and randomly selected 42 items for the mid-term exam. Next, students conducted a rudimentary item analysis of responses to evaluate the quality of the test.

Reliability from empirical data equaled 0.73, quite high for a classroom test. Eight items contained relatively high point-biserial correlations (> .50).

Literature Review

Theoreticians and practicing teachers frequently agree that classroom tests, most of which are criterion-referenced, should assess students' understanding of key concepts for a given unit of instruction. While this may be true, an assessment containing more error variance than true score variance cannot accurately estimate a student's true ability.

A considerable amount of research has been conducted on classroom tests. Within this realm, the most popular categories include studies on reliability and validity (Griswold, 1990; Mertler, 1999), guidelines for writing better tests (Fitt, Rafferty, Presner, and Heverly, 1999; Long, 1982; Thompson, Beckmann, and Senk, 1997), and techniques of test construction (O'Brien and Hampilos, 1984; Mills, 1998; Gentry, 1989; Griswold, 1990; Ornstein and Gilman, 1991). Almost all studies focus on tests made by teachers. Only Odafe (1998) has written about test construction by students for students. However, he did not analyze data generated from those tests.

Since most classroom tests have poor reliability, usually between 0.40 and 0.50, researchers tend to question the validity of these tests. If a test is not reliable, then it cannot be valid. Long (1982) asserts teachers fail to follow basic design techniques when writing tests for their classes, thereby yielding unreliable tests with little or no validity. Scores of articles offer guidelines to write classroom tests that are more psychometrically informative. While such guidelines are useful, Boothroyd (1990) points out that "about 40% of secondary teachers lack the level of measurement competency … necessary to develop effective classroom tests" (2355). He attributes teachers' "lack of measurement knowledge … to inadequate measurement training given that 51 percent of the teachers had never taken a measurement course" (2355).

Bootstrapping selected item statistics

Classical test theory (Allen and Yen, 1979) states a student's observed score equals his true score plus random error score (X = T + E). As reliability for a test increases, E decreases. As E decreases, then X more accurately estimates T. Since reliability is a function of items' and total variances, increased variance will increase reliability. A test's reliability tends to differ among groups of examinees. Other statistics used in classical test theory include the point-biserial correlation and item difficulty, both of which are frequently used in test construction. The point-biserial, an estimate of an item's discriminatory power, measures the correlation between an item and the total score. Point-biserial correlations of 0.25 or greater are usually considered acceptable. Classical test theory defines item difficulty, as the proportion of students who answer a given item correctly. Difficulty increases as the proportion decreases. If the average score is 70%, then the average item difficulty is 0.700.

Until recently, replicating statistical experiments or studies were often prohibitively expensive. However, increasingly powerful computers have thrust a new technique, bootstrapping, into the forefront of statistical research. Bootstrapping, first described by Efron (1982), randomly samples empirical data with or without replacement to generate point and interval estimates. Other uses include Monte Carlo studies, regression analyses, and goodness-of-fit tests. Refer to Davison and Hinkley (1997) for a thorough discussion of bootstrapping techniques.

Bootstrapping a normal distribution of sample size *n* relies upon a sample mean and standard deviation to construct a sampling distribution. For a simple non-parametric bootstrap, van der Vaardt (1998) recommends generating an empirical sampling distribution by "resampling with replacement from the set $\{X_1,\ldots,X_n\}$ of original observations" (328).

The researcher then estimates $\hat{\theta}^*$, the statistic of interest, based upon the empirical sampling distribution. From here the researcher can perform statistical tests and construct confidence intervals around $\hat{\theta}^*$.

Methods

Methods used in this project were similar to those described by Odafe (1998). Twenty-one students (n = 21) enrolled in a graduate-level educational assessment course were randomly assigned to one of seven groups. Each group was then randomly assigned one chapter from the course text. Groups were instructed to write fifteen questions from their respective chapters for use on the course's mid-term exam. Students could write items collaboratively or individually within their respective groups but were prohibited from collaborating with students from other groups. Students could use any combination of three item formats – true-false, multiple-choice, or short-answer – when writing items.

A total of 105 items were submitted (36 true-false, 43 multiple-choice, and 26 short-answer). Six items were deleted from the pool because they did not meet certain criteria described in the instructions to the class. Forty-two items, 17 true-false, 16 multiple-choice, and 9 short-answer, were randomly selected from the pool of 99 remaining items and administered to the class in the form of a mid-term examination.

After scoring the tests, nine items, 3 true-false, 5 multiple-choice, and a short-answer, were deleted from analysis because they had no variance. Classical item analysis was conducted on the remaining 33 items (p = 33). Next, in accordance with van der Vaardt's 1998 guidelines, I wrote three programs. These programs used Resampling Stats® to resample with replacement 2500 iterations of the empirical data to test overall reliability (KR-20), levels of difficulty (DIFF) defined as proportion of students answering an item correctly, and point-biserial correlations (PBIS). The computer programs tested one of the three following hypotheses:

$H1_0$: The distribution function of the empirical KR-20 equals a randomly permuted distribution function with a mean of 0.500.

$H1_A$: The distribution function of the empirical KR-20 equals a randomly permuted distribution function with a mean not equal to 0.500.

The first hypothesis assumes a normal distribution of (0.5, 0.1).

$H2_0$: The distribution function of the empirical DIFF equals a randomly permuted distribution function with a mean of 0.700.

$H2_A$: The distribution function of the empirical DIFF equals a randomly permuted distribution function with a mean not equal to 0.700.

The second hypothesis assumes a normal distribution of (0.70, 0.14).

$H3_0$: The distribution function of the empirical PBIS equals a randomly permuted distribution function with a mean of 0.250.

$H3_A$: The distribution function of the empirical PBIS equals a randomly permuted distribution function with a mean not equal to 0.250.

The third hypothesis assumes a normal distribution of (0.25, 0.05).

Results

Of the 33 items analyzed, 29 (87.88%) assessed learning at Bloom's taxonomic level of knowledge. The highest level of learning assessed was application (1 item).

The first resampling program tested the following hypothesis:

$H1_0$: The distribution function of the empirical KR-20 equals a randomly permuted distribution function with a mean of 0.500.

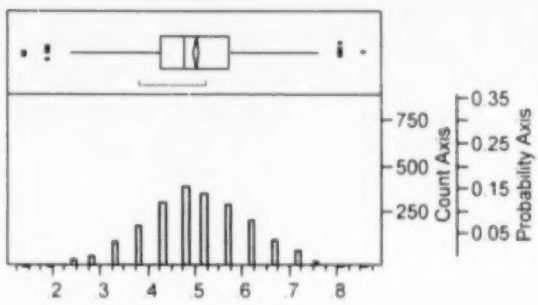
$H1_A$: The distribution function of the empirical KR-20 equals a randomly permuted distribution function with a mean not equal to 0.500.

Table 1 displays point and interval estimates for each of the three randomly permuted distribution functions. Assuming reliability for a typical classroom test is 0.500, the critical value for test the first null hypothesis was 0.500. Random permutations yielded a mean KR-20 of 0.500. The probability of observing a KR-20 greater than or equal to 0.733, given random permutations, was 0.012 (see Figure 1, page 9), which is unusual given random permutations of the data. Therefore, the data provide sufficient evidence to reject $H1_0$, implying distribution functions for empirical and resampled KR-20 are not equal.

Table 1 Point and interval estimates

Statistic	Observed	Resampled	95% CI (LL, UL)	Prob. > Observed
KR-20	.733	.499	0.286, 0.714	0.012
DIFF	.769	.702	0.476, 0.905	0.212
PBIS	.302	.251	0.095, 0.429	0.273

Figure 1: Boxplot and Histogram for distribution function of KR-20



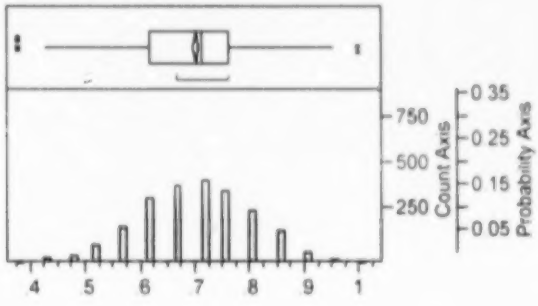
A second program tested the following hypothesis:

$H2_0$: The distribution function of the empirical DIFF equals a randomly permuted distribution function with a mean of 0.700.

$H2_A$: The distribution function of the empirical DIFF equals a randomly permuted distribution function with a mean not equal to 0.700.

The observed value for DIFF was 0.769. The second computer program applied sampling with replacement to test $H2_0$ by counting all values for DIFF greater than or equal to 0.769, then dividing this count by 2500 to derive a probability value (see Table 1 above). Random permutations of DIFF yielded a mean equal to 0700. The probability of observing a mean DIFF greater than or equal to 0.769, given random permutations, was 0.212 (see Figure 2, page 10), which is not unusual given random permutations of the data. Therefore, the data provide insufficient evidence to reject $H2_0$. Failure to reject $H2_0$ implies the distribution functions for empirical and resampled DIFF are equal.

Figure 2: Boxplot and Histogram for distribution function of DIFF



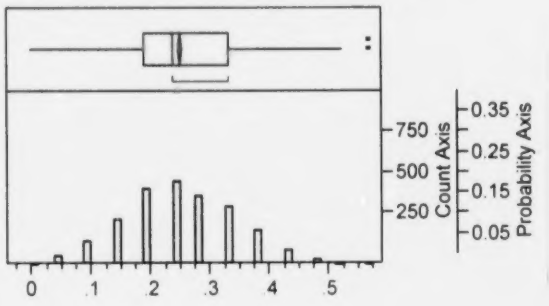
A third program tested the following hypothesis:

$H3_0$: The distribution function of the empirical PBIS equals a randomly permuted distribution function with a mean of 0.250.

$H3_A$: The distribution function of the empirical PBIS equals a randomly permuted distribution function with a mean not equal to 0.250.

The observed mean value for PBIS was 0.302. The second computer program applied sampling with replacement to test $H3_0$ by counting all values for PBIS greater than or equal to 0.302, then dividing this count by 2500 to derive a probability value (see Table 1 above). Random permutations of PBIS yielded a mean equal to 0251. The probability of observing a mean PBIS greater than or equal to 0.302, given random permutations, was 0.273 (see Figure 3, page 11), which is not unusual given random permutations of the data. Therefore, the data provide insufficient evidence to reject $H3_0$. Failure to reject $H3_0$ implies the distribution functions for empirical and resampled PBIS are equal.

Figure 3: Boxplot and Histogram for distribution function of PBIS



Discussion

The primary goal of this exercise was to demonstrate that students could write a reliable and discriminatory classroom test. Poor estimates of reliability, difficulty, and discrimination for classroom tests are not important, some educators assert, because such tests are criterion-referenced, not norm-referenced. This assertion is grounded in the idea that, for classroom purposes, understanding content is more important than identifying and ranking students based on some definition of academic ability. Such an argument is fallacious and even disingenuous.

Classroom tests, even though they are frequently criterion-referenced, still ought to discriminate the academically strong from the academically weak. Anyone can write a set of questions and corresponding options and call it a test. However, constructing a reliable classroom test that contains good discriminating properties is much more difficult.

Reliability is a function of item and total score variances; therefore, increasing variance across items and total scores will also increase reliability. To increase reliability, items must have a proper mix of difficulty levels. Items also must discriminate between ability groups. Poorly discriminating items provide no useful information and probably ought to be eliminated. Classroom tests typically have reliability estimates around 0.50, which is nothing more than noise in the system. If a typical classroom test detects nothing but noise, then the test itself is not reliable, and if a test is not reliable, then it cannot be valid.

Estimating item statistics derived from small samples is difficult. Samples are not truly random and estimates are frequently unstable. Administering a classroom test to all students enrolled in a particular subject or course is logistically difficult because teachers

cover material at different rates and order. Fortunately, the emergence of bootstrapping addresses problems associated with randomness and sample size.

Bootstrapping has several advantages. First, it permits a researcher to generate a hypothetical population by using an empirical dataset as a proxy. Second, bootstrapping minimizes bias through sampling with replacement. Third, bootstrapping is a nonparametric technique that does not rely on mathematical derivations or tables. A bootstrapped 95% confidence interval, for example, uses the 2.5 and 97.5 percentiles as the lower and upper bounds respectively. Probability values are defined as the number of observed events divided by the number of permutations.

Random permutations of KR-20, DIFF, and PBIS yielded high probability values, thereby offering insufficient evidence to reject any of the three null hypotheses. For a bootstrapped mean KR-20 of 0.500, the lower and upper percentiles of a 95% confidence interval for were 0.286 and 0.714 respectively. Reliability for this classroom test was unusual given random permutations ($p = 0.012$), and therefore estimated students' true abilities better than expected.

For a bootstrapped mean DIFF of 0.700, the lower and upper percentiles of a 95% confidence interval were 0.476 and 0.905 respectively. This confidence interval captured the empirical mean DIFF of 0.769, thereby suggesting the empirical and resampled DIFF distribution functions are equal. For a bootstrapped mean PBIS of 0.250, the lower and upper percentiles of a 95% confidence interval were 0.095 and 0.429 respectively. This confidence interval captured the empirical mean PBIS of 0.302, thereby suggesting the empirical and resampled PBIS distribution functions are equal.

Conclusion

Bootstrapping allows a researcher to test a null hypothesis against a randomly permuted distribution function. Means, standard deviations, and confidence intervals are the most commonly bootstrapped statistics. As this paper demonstrated, one can apply bootstrapping to other statistics such as KR-20, item difficulty, and discrimination.

This study has shown that creating a reliable classroom test is feasible. As such, a teacher gains more useful information about students' understanding of content. A teacher can apply this information to lesson plans, classroom instruction, and remediation.

References

Allen, M. J., and Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Company.

Boothroyd, R. A. (1990). Variables related to the characteristics and quality of classroom tests: an exploratory study with seventh- and eight-grade science and mathematics teachers. Dissertation Abstracts International, 54(07A), 2355.

Davison, A. C., and Hinkley, D. V. (1997). Bootstrap Methods and their Applications. Cambridge: Cambridge University Press.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics.

Fitt, D. X., Rafferty, K., Presner, M. T., and Heverly, M. A. (1999). Improving the quality of teachers' classroom tests. Education, 119, 643-648.

Gentry, D. L. (1989, November). Teacher-made test construction. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, AR.

Griswold, P. A. (1990). Assessing relevance and reliability to improve the quality of teacher-made tests. NASSP Bulletin, 74, 18-24.

Long, L. (1982). Writing an effective arithmetic test. Arithmetic Teacher, 29, 16-18.

Mertler, C. A. (1999, October). Teachers' (Mis)Conceptions of Classroom Test Validity and Reliability. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago, IL.

Mills, R. W. (1998, April). Development of program and individual student evaluation models for foreign language in the elementary school. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

O'Brien, M. and Hampilos, J. P. (1984, April). The feasibility of creating an item bank from a teacher-made test using the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Odafe, V. U. (1998). Students generating test items: a teaching and assessment strategy. Mathematic Teacher, 91, 198-202.

Ornstein, A. C., and Gilman, D. A. (1991). The striking contrast between norm-referenced and criterion-referenced tests. Contemporary Education, 62, 287-293.

Thompson, D. R., Beckmann, C. E., and Senk, S. L. (1997). Improving classroom tests as a means of improving assessment. Mathematics Teacher, 90, 58-64.

Vaardt, A. W. van der. (1998). Asymptotic Statistics. Cambridge: Cambridge University Press.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org